

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平9-319394

(43)公開日 平成9年(1997)12月12日

(51)Int.Cl.⁶
G 10 L 5/04

識別記号 庁内整理番号

F I
G 10 L 5/04技術表示箇所
E
A
F
H

3/00

3/00

審査請求 未請求 請求項の数9 OL (全17頁)

(21)出願番号

特願平9-46694

(22)出願日

平成9年(1997)2月28日

(31)優先権主張番号 特願平8-54714

(32)優先日 平8(1996)3月12日

(33)優先権主張国 日本 (JP)

(31)優先権主張番号 特願平8-77393

(32)優先日 平8(1996)3月29日

(33)優先権主張国 日本 (JP)

(71)出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72)発明者 笠嶋 岳彦

神奈川県川崎市幸区小向東芝町1番地 株式会社東芝研究開発センター内

(72)発明者 赤嶺 政巳

神奈川県川崎市幸区小向東芝町1番地 株式会社東芝研究開発センター内

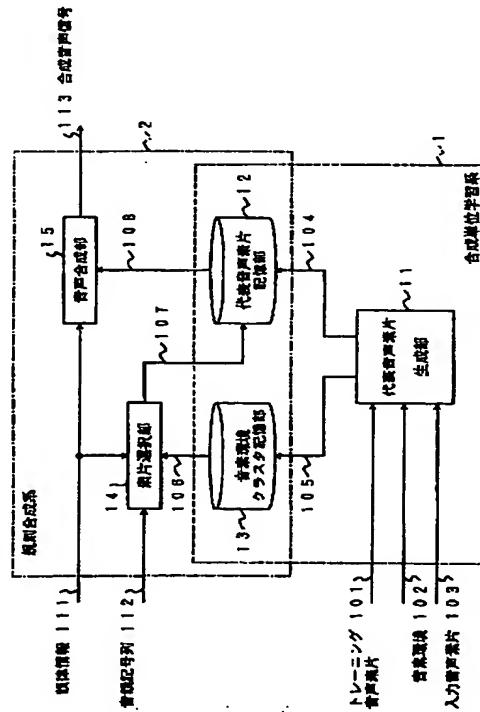
(74)代理人 弁理士 鈴江 武彦 (外6名)

(54)【発明の名称】 音声合成方法

(57)【要約】

【課題】テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供する。

【解決手段】代表音声素片生成部11において音素環境102がラベル付けされたトレーニング音声素片101のピッチ・継続時間長に従って入力音声素片103のピッチ・継続時間長を変更して複数の合成音声素片を生成し、合成音声素片とトレーニング音声素片101との間の距離尺度に基づいて入力音声素片103から代表音声素片104を選択して代表音声素片記憶部12に記憶し、さらに距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタ105を生成して音素環境クラスタ記憶部13に記憶し、代表音声素片記憶部12から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を読み出して音声合成部15で接続することにより、合成音声信号113を生成する。



し、これらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と前記第1の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、

前記距離尺度に基づいて前記第2の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、

これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成し、

この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項9】複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って代表音声素片を用いて複数の合成音声素片を生成し、

これらの合成音声素片と前記複数の第1の音声素片との間で定義される歪みの評価関数に基づいて複数の前記代表音声素片を求めて記憶し、

これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、テキスト音声合成のための音声合成方法に係り、特に音韻記号列、ピッチおよび音韻継続時間長などの情報から音声信号を生成する音声合成方法に関する。

【0002】

【従来の技術】任意の文章から人工的に音声信号を作り出すことをテキスト音声合成という。テキスト音声合成は、一般的に言語処理部、音韻処理部および音声合成部の3つの段階によって行われる。入力されたテキストは、まず言語処理部において形態素解析や構文解析などが行われ、次に音韻処理部においてアクセントやイントネーションの処理が行われて、音韻記号列・ピッチ・音韻継続時間長などの情報がoutputされる。最後に、音声信号合成部で音韻記号列・ピッチ・音韻継続時間長などの情報から音声信号を合成する。そこで、テキスト音声合成に用いる音声合成方法は、任意の音韻記号列を任意の韻律で音声合成することが可能な方法でなければならない。

【0003】このような任意の音韻記号列を音声合成する音声合成装置の原理は、母音をV、子音をCで表すと、CV、CVC、VCVといった基本となる小さな単位の特徴パラメータ（これを代表音声素片という）を記憶し、これらを選択的に読み出した後、ピッチや継続時間長を制御して接続することにより、音声を合成するというものである。従って、記憶されている代表音声素片が合成音声の品質を大きく左右することになる。

【0004】従来、これらの代表音声素片の作成はもっぱら人手に頼っており、音声信号の中から試行錯誤的に切り出していく場合がほとんどであるため、膨大な労力を要していた。このような代表音声素片作成の作業を自動化し、音声合成に使用するのに適した代表音声素片を容易に生成する方法として、例えば音素環境クラスタリング(COC)と呼ばれる技術が特開昭64-78300「音声合成方法」に開示されている。

【0005】COCの原理は、音素名や音素環境のラベルを多数の音声素片に付与し、そのラベルが付与された音声素片を音声素片間の距離尺度に基づいて音素環境に関する複数のクラスタに分類し、その各クラスタのセントロイドを代表音声素片とするものである。ここで、音素環境とは当該音声素片にとっての環境となる要因全ての組合せであり、その要因としては当該音声素片の音素名、先行音素、後続音素、後々続音素、ピッチ周期、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情などが考えられる。実音声中の各音素は音素環境によって音韻が変化しているため、音素環境に関する複数のクラスタ毎に代表素片を記憶しておくことにより、音素環境の影響を考慮した自然な音声を合成することが可能となっている。

【0006】

【発明が解決しようとする課題】上に述べたように、テキスト音声合成のための音声合成では、代表音声素片のピッチや継続時間長を指定された値に変更して合成する必要がある。このようなピッチや継続時間長の変更により、代表音声素片を切り出してきた音声信号の音質と比較して合成音声の音質がある程度劣化することになる。

【0007】これに対して、上記のCOCによるクラスタリングでは、音声素片間の距離尺度に基づいてクラスタリングを行っているにすぎないため、合成際のピッチや継続時間の変更の効果が全く考慮されていないという問題がある。すなわち、COCによるクラスタリングおよび各クラスタの代表音声素片は、実際にピッチや継続時間長を変更して合成された合成音声のレベルでは、必ずしも適当なものになっているという保証はない。

【0008】本発明は、このような問題点を解決すべくなされたものであり、テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供することを目的とする。

【0009】

【課題を解決するための手段】上記の課題を解決するため、本発明はピッチや継続時間長の変更の影響を考慮して、合成音声のレベルで自然音声に対する歪みが小さくなるような代表音声素片を生成し、その代表音声素片を用いて音声を合成することにより、自然音声に近い合成音声を生成するようにしたものである。

【0010】すなわち、本発明に係る音声合成方法は、複数の第1の音声素片のピッチおよび継続時間長の少な

くとも一方で複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて第2の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする。

【0011】ここで、第1および第2の音声素片は、C V, VCV, CVCといった音声合成単位で音声信号中から切り出される素片であり、切り出された波形もしくはその波形から何らかの方法で抽出されたパラメータ系列などを表すものとする。これらのうち、第1の音声素片は合成音声の歪みを評価するために用いられ、また第2の音声素片は代表音声素片の候補として用いられる。合成音声素片は、第2の音声素片に対して少なくともピッチまたは継続時間長を変更して生成される合成音声波形またはパラメータ系列などを表す。

【0012】合成音声素片と第1の音声素片との間の距離尺度によって、合成音声の歪みが表わされる。従つて、この距離尺度つまり歪みがより小さくなる音声素片を第2の音声素片から選択して代表音声素片として記憶しておき、これらの代表音声素片から所定の代表音声素片を選択して接続すれば、自然音声に近い高品質の合成音声が生成される。

【0013】本発明の第1の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方で複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて第2の音声素片から複数の代表音声素片を選択して記憶し、前記距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、複数の代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【0014】ここで、音素環境とは前述した通り音声素片にとっての環境となる要因、例えば当該音声素片の音素名、先行音素、後続音素、後々続音素、ピッチ周期、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情といった要素の組み合わせであり、音素環境クラスタとは言い換えれば音素環境の集合であり、例えば「当該素片の音韻が/k a/、先行音韻が/i/または/u/、ピッチ周波数が200Hz以下」というようなものを意味する。

【0015】第1の態様のように、距離尺度つまり合成音声の歪みに基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続するようにすれば、例えば同一音素名の音声素

片が複数の音素環境に存在する場合でも、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【0016】本発明の第2の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方で複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて第2の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成する。この第2の態様は、音声素片が一つの音素環境にのみ存在する場合に有効である。

【0017】本発明の第3の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方で複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、第1の音声素片と合成音声素片との間の距離尺度に基づいて第2の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【0018】この第3の態様によっても、第1の態様と同様に、例えば同一音素名の音声素片が複数の音素環境に存在する場合、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【0019】また、本発明に係る他の音声合成方法は、複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方で複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、さらにこれらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と第1の音声素片との間の距離尺度に基づいて第2の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする。

【0020】この場合、先に示した第1、第2および第3の態様においても、複数の合成音声素片を生成した後、スペクトル整形を行うようにする。ここで、スペクトル整形は「めりはり」のある明瞭な音声を合成するための処理であり、例えばホルマント強調やピッチ強調を

行う適応ポストフィルタによるフィルタリングによって実現される。

【0021】このように代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整形を行うことによって、スペクトル整形後の最終的な合成音声のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」に優れたより明瞭な合成音声が得られる。

【0022】本発明に係るさらに別の音声符号化方法は、複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って代表音声素片を用いて複数の合成音声素片を生成し、これらの合成音声素片と複数の第1の音声素片との間で定義される歪みの評価関数に基づいて複数の代表音声素片を求めて記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする。

【0023】このように第2に音声素片から代表音声素片を選択するのではなく、第2の音声素片を用いずに第1の音声素片に対して最適な代表音声素片を生成することも可能である。このようにして生成された代表音声素片から、先と同様に所定の代表音声素片を選択して接続することにより、自然音声に近い高品質の合成音声が生成される。

【0024】

【発明の実施の形態】以下、図面を参照して本発明の一実施形態を説明する。図1は、本発明の一実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図であり、大きく分けて合成単位学習系1と規則合成系2からなる。実際にテキスト音声合成を行う場合に動作するのは規則合成系2であり、合成単位学習系1は事前に学習を行って代表音声素片を生成するものである。

【0025】まず、合成単位学習系1について説明する。合成単位学習系1は、代表音声素片とこれに付随する音素環境クラスタを生成する代表音声素片生成部11と代表音声素片記憶部12および音素環境クラスタ記憶部13により構成される。代表音声素片生成部11には、第1の音声素片であるトレーニング音声素片101とこれにラベル付けされた音素環境102および第2の音声素片である入力音声素片103が入力される。

【0026】代表音声素片生成部11では、トレーニング音声素片101にラベル付けされた音素環境102に含まれるピッチ周期および継続時間長の情報に従って、入力音声素片103のピッチ周期および継続時間長を変更することで複数の合成音声素片が内部的に生成され、さらにこれらの合成音声素片とトレーニング音声素片101との距離尺度に従って、代表音声素片104と音素環境クラスタ105が生成される。音素環境クラスタ105は、トレーニング音声素片101を後述するように

音素環境に関するクラスタに分類して生成される。

【0027】代表音声素片104は代表音声素片記憶部12に記憶され、音素環境クラスタ105は代表音声素片104と対応付けられて音素環境クラスタ記憶部13に記憶される。代表音声素片生成部11の処理については、後に詳細に説明する。

【0028】次に、規則合成系2について説明する。規則合成系2は、代表音声素片記憶部12と音素環境クラスタ記憶部13と素片選択部14および音声合成部15により構成され、代表音声素片記憶部12と音素環境クラスタ記憶部13を合成単位学習系1と共有している。

【0029】素片選択部14には、入力音素の情報として、例えばテキスト音声合成のために入力テキストの形態素解析・構文解析後さらにアクセントやイントネーション処理を行って得られた韻律情報111と音韻記号列112が入力される。韻律情報111には、ピッチパターンおよび音韻継続時間長が含まれている。素片選択部14では、これらの韻律情報111と音韻記号列112から入力音素の音素環境を内部的に生成する。

【0030】そして、素片選択部14は音素環境クラスタ記憶部13より読み出された音素環境クラスタ106を参照して、入力音素の音素環境がどの音素環境クラスタに属するかを探索し、探索した音素環境クラスタに対応する代表音声素片選択情報107を代表音声素片記憶部12へ出力する。

【0031】音声合成部15は、代表音声素片選択情報107に従って代表音声素片記憶部12より選択的に読み出された代表音声素片108に対して、韻律情報111に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号113を出力する。ここで、ピッチおよび継続時間長を変更して素片を接続し音声を合成する方法としては、例えば残差駆動LSP方法や波形編集方法など公知の技術を用いることができる。

【0032】次に、本発明の特徴をなす代表音声素片生成部11の処理の実施形態について具体的に説明する。図2のフローチャートは、代表音声素片生成部11の第1の実施形態による処理手順を示している。

【0033】この第1の実施形態による代表音声素片生成処理では、まず準備段階として連続発声された多数の音声データに対して音韻毎にラベリングを行い、CV, VCV, CVCなどの合成単位に従って、トレーニング音声素片 T_i ($i = 1, 2, 3, \dots, N_T$) を切り出す。また、各トレーニング音声素片 T_i に対応する音素環境 P_i ($i = 1, 2, 3, \dots, N_T$) も抽出しておく。ただし、 N_T はトレーニング音声素片の個数を表す。音素環境 P_i は、少なくともトレーニング音声素片 T_i の音韻とそのピッチおよび継続時間長の情報を含むものとし、その他に必要に応じて前後の音素などの情報を含むものとする。

【0034】次に、上述したトレーニング音声素片 T_i の作成と同様の方法により、多数の入力音声素片 S_i ($i = 1, 2, 3, \dots, N_s$) を作成する。ただし、 N_s は入力音声素片の個数を表す。ここで、入力音声素片 S_i としてはトレーニング音声素片 T_i と同じものを使用してもよいし（すなわち $T_i = S_i$ ）、トレーニング音声素片 T_i とは異なる音声素片を作成してもよい。いずれにしても、豊富な音韻環境を有する多数のトレーニング音声素片および入力音声素片が用意されていることが望ましい。

【0035】このような準備段階を経た後、まず音声合成ステップ S21 で、音素環境 P_i に含まれるピッチおよび継続時間長に等しくなるように、入力音声素片 S_i のピッチおよび継続時間長を変更して音声を合成することにより、合成音声素片 G

i_j を生成する。ここでのピッチおよび継続時間長の変更は、音声合成部 15 におけるピッチおよび継続時間長の変更と同様の方法で行われるものとする。全ての音素環境 P_i ($i = 1, 2, 3, \dots, N_t$) に従って入力音声素片 S_i ($j = 1, 2, 3, \dots, N_s$) を用いて音声の合成を行うことにより、 $N_t \times N_s$ 個の合成音声素片 G_{ij} ($i = 1, 2, 3, \dots, N_t$, $j = 1, 2, 3, \dots, N_s$) を生成する。

【0036】次に、歪み評価ステップ S22 では、合成音声素片 G_{ij} の歪み e_{ij} の評価を行う。この歪み e_{ij} の*

$$E_{ij}(U) = \sum_{i=1}^{N_t} \min(e_{ij1}, e_{ij2}, e_{ij3}, \dots, e_{ijN}) \quad (1)$$

【0040】ただし、 $\min(e_{ij1}, e_{ij2}, e_{ij3}, \dots, e_{ijN})$ は $e_{ij1}, e_{ij2}, e_{ij3}, \dots, e_{ijN}$ の中の最小値を表す関数である。集合 U の組合せは $N_s ! / (N_s - N) !$ 通りあり、これらの音声素片の集合 U の中から評価関数 $E_{ij}(U)$ を最小にする U を探索し、その要素 u_k を代表音声素片 D_k とする。

【0041】最後に、音素環境クラスタ生成ステップ S

$$E_{ci} = \sum_{k=1}^N \sum_{P_i \in C_k} e_{ijk}$$

【0043】こうしてステップ S23 および S24 で生成された代表音声素片 D_k および音素環境クラスタ C_k は、図 1 の代表音声素片記憶部 12 および音素環境クラスタ記憶部 13 にそれぞれ記憶される。

【0044】次に、図 3 のフローチャートを参照して代表音声素片生成部 11 の第 2 の実施形態による処理手順について説明する。この第 2 の実施形態による代表音声素片生成処理では、まず初期音素環境クラスタ生成ステップ S30 において、何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。音素環境のクラスタリングには、例

*評価は、合成音声素片 G_{ij} とトレーニング音声素片 T_i との間の距離尺度を求めるこにより行う。距離尺度には、何らかのスペクトル距離を用いることができる。例えば、合成音声素片 G_{ij} およびトレーニング音声素片 T_i について、FFT (高速フーリエ変換) などを用いてパワースペクトルを求めて各パワースペクトル間の距離を評価する方法や、あるいは線形予測分析を行って LPC または LSP パラメータなどを求めて各パラメータ間の距離を評価する方法などがある。その他にも、短時間

10 フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法も用いることができる。また、各素片のパワーを正規化した上で歪みの評価を行う方法でもよい。

【0037】次に、代表音声素片生成ステップ S23 では、ステップ S22 で得られた歪み e_{ij} に基づいて、入力音声素片 S_i の中から指定された代表音声素片数 N の代表音声素片 D_k ($k = 1, 2, 3, \dots, N$) を選択する。

【0038】代表音声素片選択法の一例を説明する。入力音声素片 S_i の中から選択された N 個の音声素片の集合 $U = \{u_k \mid u_k = S_i \ (k = 1, 2, 3, \dots, N)\}$ に対して、歪みの総和を表す評価関数 $E_{ij}(U)$ を次式 (1) のように定義する。

【0039】

【数1】

【0040】ただし、音素環境 P_i 、歪み e_{ij} および代表音声素片 D_k より、音素環境に関する複数のクラスタ（音素環境クラスタ） C_k ($k = 1, 2, 3, \dots, N$) を生成する。音素環境クラスタ C_k は、例えば次式 (2) で表されるクラスタリングの評価関数 E_{ci} を最小化するクラスタを探索することによって得られる。

【0042】

【数2】

(2)

【0043】そして音韻によるクラスタリングを行うことができる。

【0044】そして、入力音声素片 S_i およびトレーニング音声素片 T_i のうち音韻が一致する音声素片のみをそれぞれ用いて、図 2 のステップ S21, S22, S23, S24 と同様の合成音声素片生成ステップ S31、歪み評価ステップ S32、代表音声素片生成ステップ S33、音素環境クラスタ生成ステップ S34 の処理を順次行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことにより、全ての代表音声素片およびそれに対応する音素環境クラスタの生成を行う。こうして生成された代表音声素片および音素環境クラスタは、図

1の代表音声素片記憶部12および音素環境クラスタ記憶部13にそれぞれ記憶される。

【0046】ただし、各初期音素環境クラスタ当たりの代表音声素片数が1であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、音素環境クラスタ生成ステップS34は不要となり、初期音素環境クラスタを音素環境クラスタ記憶部13に記憶すればよい。

【0047】次に、図4のフローチャートを参照して代表音声素片生成部11の第3の実施形態による処理手順を説明する。この第3の実施形態による代表音声素片生

$$E_{C2} = \sum_{k=1}^N \min\{f(k,1), f(k,2), f(k,3), \dots, f(k,N)\} \quad (3)$$

$$f(k,j) = \sum_{p_i \in C_k} e_{i,j} \quad (4)$$

【0049】次に、代表音声素片生成ステップS44において、歪み $e_{i,j}$ に基づいて音素環境クラスタ C_k のそれぞれに対応する代表音声素片 D_k を入力音声素片 S_j より選択する。この代表音声素片 D_k は、入力音声素片 S_j から例えば次式(5)で表される歪み評価関数 E_{D2} ※

$$E_{D2}(j) = \sum_{p_i \in C_k} e_{i,j} \quad (5)$$

【0051】なお、この第3の実施形態による代表音声素片生成処理を変形し、第2の実施形態と同様に、何らかの先見的な知識に基づいて予め生成した初期音素環境クラスタ毎に代表音声素片の生成および音素環境クラスタの生成を行うことも可能である。

【0052】次に、図5～図9を用いて本発明の他の実施形態について説明する。図5は、本発明の他の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。図1と相対応する部分に同一の参照符号を付して相違点を中心に説明すると、本実施形態では音声合成部15の後段に適応ポストフィルタ16が追加されている点が先の実施形態と異なり、これに加えて代表音声素片生成部11における複数の合成音声素片の生成法も先の実施形態と異なっている。

【0053】すなわち、代表音声素片生成部11では先の実施形態と同様に、トレーニング音声素片101にラベル付けされた音素環境102に含まれるピッチ周期および継続時間長の情報を従って、入力音声素片103のピッチ周期および継続時間長を変更することで複数の合成音声素片を内部的に生成した後、これらの合成音声素片に対して適応ポストフィルタによるフィルタリングを施してスペクトル整形を行う。そして、この適応ポストフィルタによりスペクトル整形を行った後の各合成音声素片とトレーニング音声素片101との距離尺度に従って、代表音声素片104と音素環境クラスタ105が生成される。音素環境クラスタ105は、先の実施形態と同様にトレーニング音声素片101を音素環境に関する

*成処理では、図2に示した第1の実施形態と同様に音声合成ステップS41および歪み評価ステップS42を順次経た後、次の音素環境クラスタ生成ステップS43において、音素環境 P_i および歪み $e_{i,j}$ に基づいて音素環境に関するクラスタ C_k ($k = 1, 2, 3, \dots, N$) を生成する。音素環境クラスタ C_k は、例えば次式(3)

(4)で表わされるクラスタリングの評価関数 E_{C2} を最小化するクラスタを探索することによって得られる。

【0048】

10 【数3】

※(j)を最小化する音声素片を探索することによって得られる。

【0050】

20 【数4】

クラスタに分類して生成される。

【0054】なお、この代表音声素片生成部11において音素環境102に含まれるピッチ周期および継続時間長の情報を従って入力音声素片103のピッチ周期および継続時間長を変更して生成される複数の合成音声素片に対してフィルタリングを施してスペクトル整形を行う適応ポストフィルタは、音声合成部15の後段に配置される適応ポストフィルタ16と同様の構成でよい。

【0055】一方、音声合成部15では先の実施形態と同様に代表音声素片選択情報107に従って代表音声素片記憶部12より選択的に読み出された代表音声素片108に対し、韻律情報111に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号113を生成するが、本実施形態ではこの合成音声信号113がさらに適応ポストフィルタ16に入力され、ここで音質向上のためのスペクトル整形が行われた後、最終的な合成音声信号114が取り出される。

【0056】図6に、適応ポストフィルタ16の一構成例を示す。この適応ポストフィルタ16は、ホルマント強調フィルタ21とピッチ強調フィルタ22を継続配置して構成される。

【0057】ホルマント強調フィルタ21は、代表音声素片選択情報107に従って代表音声素片記憶部12から選択的に読み出された代表音声素片108をLPC分析して得られるLPC係数に基づいて決定されるフィルタ係数に従って、音声合成部15から入力される合成音

声信号113をフィルタリングすることにより、スペクトルの山の部分を強調する処理を行う。一方、ピッチ強調フィルタ22は、韻律情報111に含まれるピッチ周期に基づいて決定されるパラメータに従って、ホルマント強調フィルタ21の出力をフィルタリングすることにより、音声信号のピッチを強調する処理を行う。なお、ホルマント強調フィルタ21とピッチ強調フィルタ22の配置順序は逆であってもよい。

【0058】このような適応ポストフィルタ16の適用によりスペクトルが整形され、「めりはり」のある明瞭な音声を再生可能な合成音声信号114が得られる。適応ポストフィルタ16としては図6に示した構成のものに限られず、音声符号化や音声合成の分野で用いられる公知の技術に基づく種々の構成を採用することが可能である。

【0059】このように本実施形態では、規則合成系2において音声合成部15の後段に適応ポストフィルタ16が配置される点を考慮して、合成単位学習系1においても代表音声素片生成部11で音素環境102に含まれるピッチ周期および維続時間長の情報に従って入力音声素片103のピッチ周期および維続時間長を変更して生成される複数の合成音声素片に対し、同様に適応ポストフィルタによるフィルタリングを行っている。従って、適応ポストフィルタ16を通した後の最終的な合成音声信号114と同様のレベルで、自然音声に対する歪みが小さくなるような代表音声素片を代表音声素片生成部11において生成できるため、さらに自然音声に近い合成音声を生成することが可能となる。

【0060】次に、図5における代表音声素片生成部11の処理の実施形態について具体的に説明する。図7、図8および図9のフローチャートは、図5における代表音声素片生成部11の第1、第2および第3の実施形態による処理手順を示している。図7、図8および図9では、先に説明した図2、図3および図4に示した処理手順における音声合成ステップS21、S31およびS41の後に、ポストフィルタリングステップS25、S36およびS45が追加されている。

【0061】ポストフィルタリングステップS25、S36およびS45では、前述した適応ポストフィルタによるフィルタリングを行う。すなわち、音声合成ステップS21、S31およびS41で生成された合成音声素片G_iに対し、入力音声素片S_iをLPC分析して得られるLPC係数に基づいて決定されるフィルタ係数に従ってフィルタリングを行うことにより、スペクトルの山の部分を強調するホルマント強調を行う。また、このホルマント強調後の合成音声素片に対し、さらにトレーニング音声素片T_iのピッチ周期に基づいて決定されるパラメータに従ってフィルタリングを行うことにより、ピッチ強調を行う。

【0062】このようにして、ポストフィルタリングス

テップS25、S36およびS45において、スペクトル整形を行う。このポストフィルタリングステップS25、S36およびS45は、前述したように規則合成系2において音声合成部15の後段に設けられる適応ポストフィルタ16により合成音声信号113のスペクトル整形を行って音質の向上を図るポストフィルタリングを行うことを前提に、合成単位の学習を可能とする処理であり、この処理を適応ポストフィルタ16による処理と組み合わせることによって、最終的に「めりはり」のある明瞭な合成音声信号114が生成される。

【0063】次に、図10～図13を用いて本発明の別の実施形態について説明する。図10は、本発明の他の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。図1と相対応する部分に同一の参照符号を付して相違点を中心に説明すると、本実施形態では代表音声素片生成部31に入力音声素片103が入力されていない点がこれまでの実施形態と異なっている。

【0064】すなわち、本実施形態では先の実施形態のように入力音声素片103の中から選択した音声素片を代表音声素片104とするのではなく、トレーニング音声素片101に対して最適な代表音声素片104を計算によって新たに生成する。音素環境クラスタ105は、先の実施形態と同様にトレーニング音声素片101を音素環境に関するクラスタに分類して生成される。

【0065】次に、図10における代表音声素片生成部31の処理の実施形態について具体的に説明する。図11のフローチャートは、代表音声素片生成部31の第1の実施形態による処理手順を示している。この第1の実施形態による代表音声素片生成処理では、先の実施形態の代表音声素片生成部11における代表音声素片生成処理と同様に、まず、準備段階として連続発声された多数の音声データに音韻毎にラベリングを行い、CV、VCV、CVCなどの合成単位に従ってトレーニング音声素片T_i (i=1, 2, 3, ..., N_T) を切り出す。また、各トレーニング音声素片に対応する音素環境P_i (i=1, 2, 3, ..., N_P) を抽出しておく。ただし、N_Tはトレーニング音声素片の個数を表す。音素環境は、少なくとも当該トレーニング音声素片の音韻とそのピッチパターンおよび維続時間長を含むものとし、その他に必要に応じて前後の音素などを含むものとする。

【0066】このような準備段階を経た後、まず代表音声素片初期化ステップS51で、指定された代表音声素片数Nの代表音声素片D_k (k=1, 2, 3, ..., N) を初期化して初期代表音声素片D_k⁰ (k=1, 2, 3, ..., N) を生成する。初期代表音声素片D_k⁰としては、任意の音声素片を用いることが可能であり、例えばトレーニング音声素片T_iからランダムに選択された素片を用いることができる。

【0067】次に、音声合成ステップS52で、D_k⁰

のピッチおよび継続時間長を P_i のピッチパターンおよび継続時間長に等しくなるように変更して音声を合成して合成音声素片 G_{ik} を生成する。ここでのピッチおよび継続時間長の変更は、音声合成部 20 におけるピッチおよび継続時間長の変更と同様の方法で行われるものとする。全ての P_i ($i = 1, 2, 3, \dots, N_T$) に従って D_k^0 ($k = 1, 2, 3, \dots, N_T$) を用いて合成を行うことにより、 $N_T \times N$ 個の合成音声素片 G_{ik} ($i = 1, 2, 3, \dots, N_T, k = 1, 2, 3, \dots, N$) を生成する。

【0068】次に、歪み評価ステップ S 5 3 では、合成音声素片 G_{ik} とトレーニングとの間で定義される歪み e_{ij} の評価を行う。歪みの評価法としては、波形の 2 乗誤差や何らかのスペクトル距離を用いることができる。例えば、FFTなどを用いてパワースペクトルを求めてそ*

$$E_{ci} = \sum_{k=1}^N \sum_{P_i \in C_k} e_{ik}$$

【0071】但し、次式に示されるように全ての音素環境クラスタ C_k ($k = 1, 2, 3, \dots, N$) の和集合は音素環境の全体集合 A と等しく、かつ任意の異なる 2 つの音素環境クラスタの積集合は空集合 ϕ になるものとする

$$C_1 \cup C_2 \cup \dots \cup C_N = A$$

$$C_i \cap C_j = \phi \quad (i \neq j)$$

【0073】次に、代表音声素片生成ステップ S 5 5 では、代表音声素片を更新するため、クラスタ C_k に対応する代表音声素片 D_k^1 をクラスタ毎に求める。代表音声素片 D_k^1 は、クラスタに属するトレーニング音声素片と、対応する合成音声素片との歪みの総和を表す評価★30

$$ED_3(D_k) = \sum_{P_i \in C_k} |t_i - g_{ik}(D_k)|^2$$

【0075】但し、 t_i はトレーニング音声素片 T_i の波形を表すベクトル、 $g_{ik}(D_k)$ は P_i に従って D_k を用いて合成された合成音声素片 G_{ik} の波形を表すベクトルである。また、評価関数の例としては、これ以外にもパワースペクトルの距離やLPC, LSPなどのパラメータ間の距離の総和などがある。評価関数を代表音声素片で偏微分したものを 0 とおいた方程式が解ける場合には、これを解いて評価関数を最小にする代表音声素片を解析的に求めることができる。それ以外の場合は、公知の最適化手法を用いて代表音声素片を求めることができる。

【0076】このようにステップ S 5 2 から S 5 5 までの処理で、初期代表音声素片 D_k^0 から代表音声素片 D_k^1 に更新される。代表音声素片が更新されることによって音素環境クラスタが変化するため、代表音声素片および音素環境クラスタの変化が十分小さくなり収束するまで、ステップ S 5 2 から S 5 5 までの処理を繰り返す必要がある。

*の間の距離を求める方法や、あるいは線形予測分析を行ってLPCまたはLSPパラメータなどを求めてパラメータ間の距離を評価する方法などがある。その他にも、短時間フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法が考えられる。また、各素片のパワーを正規化した上で歪みの評価を行うことも考えられる。

【0069】次に、音素環境クラスタ生成ステップ S 5 4 では音素環境 P_i および歪み e_{ik} に基づいて、音素環境 10 に関するクラスタ C_k ($k = 1, 2, 3, \dots, N$) を生成する。音素環境クラスタ C_k は、例えば次式で表されるクラスタリングの評価関数 E_{c3} を最小化するクラスタを探索することによって得られる。

【0070】

【数 5】

(6)

※る。

20 【0072】

【数 6】

(7)

【0074】

【数 7】

(8)

★関数が最小になるように計算される。評価関数としては、例えば次式に示す波形の 2 乗誤差の総和を用いることができる。

【0074】

【数 7】

(9)

【0077】そこで、次の収束判定ステップ S 5 6 では、更新の前後における代表音声素片の変化の度合から代表音声素片および音素環境クラスタの変化が収束したか否かを判定し、収束していないと判定された場合はステップ S 5 2 から S 5 5 までの処理を繰り返してさらに代表音声素片を更新し、収束したと判定された場合は処理を終了し、最新の代表音声素片 D_k^m ($k = 1, 2, 3, \dots, N, m$ は繰り返し回数) が代表音声素片 D_k ($k = 1, 2, 3, \dots, N$) となる。

【0078】こうして生成された代表音声素片 D_k および音素環境クラスタ C_k は、図 10 の代表音声素片記憶部 12 および音素環境クラスタ記憶部 13 にそれぞれ記憶される。

【0079】次に、図 12 のフローチャートを参照して代表音声素片生成部 31 の第 2 の実施形態による処理手順について説明する。この第 2 の実施形態による代表音声素片生成処理では、まず初期音素環境クラスタ生成ステップ S 6 1 において、何らかの先見的な知識に基づい

て予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。音素環境のクラスタリングには、例えば音韻によるクラスタリングを行うことができる。

【0080】そして、トレーニング音声素片 T_i のうち音韻が一致する音声素片のみを用いて、図11のステップS51、S52、S53、S54、S55、S56と同様の代表音声素片初期化ステップS62、音声合成ステップS63、歪み評価ステップS64、音素環境クラスタ生成ステップS65、代表音声素片生成ステップS66、収束判定ステップS67の処理を順次行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことにより、全ての代表音声素片およびそれに対応する音素環境クラスタの生成を行う。こうして生成された代表音声素片および音素環境クラスタは、図10の代表音声素片記憶部12および音素環境クラスタ記憶部13にそれぞれ記憶される。

【0081】但し、各初期音素環境クラスタ当りの代表音声素片数が1であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、ステップS62、S63、S64、S65、S67の処理は不要とな*

$$E_{C4} = \sum_{k=1}^N \sum_{T_i \in C'_k} c_{i,k}$$

【0085】次に、図11のステップS55、S56と同様の代表音声素片生成ステップS75、収束判定ステップS76の処理を順次行って、代表音声素片およびそれに対応するトレーニング音声素片クラスタが生成される。

【0086】最後に、音素環境クラスタ生成ステップS77では、トレーニング音声素片クラスタ C'_k に属するトレーニング音声素片 T_i に共通する音素環境を抽出して音素環境クラスタ C_k を生成する。但し、音素環境クラスタ C_k ($k = 1, 2, 3, \dots, N$) は、式(7) (8) の条件を満たすものとする。また、本実施形態の音声合成方法に前の実施形態と同様にポストフィルタリング処理を組み合わせることも可能である。

【0087】

【発明の効果】以上説明したように、本発明の音声合成方法によれば、入力音声素片に対してピッチおよび継続時間長の少なくとも一方の変更を行って生成される合成音声のレベルで自然音声に対する歪みを評価し、その歪み評価結果に基づいて入力音声素片から選択した音声素片を代表音声素片とするか、あるいは歪み評価結果に基づいて代表音声素片を生成するため、音声合成装置の特性をも考慮した代表音声素片の生成が可能であり、この代表素片を接続して音声合成を行うことによって、自然音声に近い高品質の合成音声を生成することができる。

【0088】また、本発明ではさらに代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整

*り、初期音素環境クラスタに対応する代表音声素片を代表音声素片生成ステップS66で求めればよい。この場合には、初期音素環境クラスタを音素環境クラスタ記憶部13に記憶すればよい。

【0082】次に、図13のフローチャートを参照して代表音声素片生成部31の第3の実施形態による処理手順について説明する。まず、図11のステップS51、S52、S53と同様の代表音声素片初期化ステップS71、音声合成ステップS71、歪み評価ステップS73の処理を順次行って、合成音声素片 G_{ik} とトレーニング音声素片 T_i の間の歪み e_{ik} を求める。

【0083】次に、トレーニング音声素片クラスタ生成ステップS74では、歪み e_{ik} に基づいてトレーニング音声素片 T_i のクラスタ C'_k ($k = 1, 2, 3, \dots, N$) を生成する。このトレーニング音声素片クラスタ C'_k は、例えば次式で表されるクラスタリングの評価関数 E_{C4} を最小化するクラスタを探索することによって得られる。

【0084】

【数8】

形を行うことにより、スペクトル整形後の最終的な合成音声信号のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」のあるより明瞭な合成音声を生成することができる。

【図面の簡単な説明】

【図1】本発明の一実施形態に係る音声合成装置のブロック図

【図2】図1中の代表音声素片生成部での第1の実施形態による処理手順を示すフローチャート

【図3】図1中の代表音声素片生成部での第2の実施形態による処理手順を示すフローチャート

【図4】図1中の代表音声素片生成部での第3の実施形態による処理手順を示すフローチャート

【図5】本発明の他の実施形態に係る音声合成装置のブロック図

【図6】図5中の適応ポストフィルタの構成例を示すブロック図

【図7】図5中の代表音声素片生成部での第1の実施形態による処理手順を示すフローチャート

【図8】図5中の代表音声素片生成部での第2の実施形態による処理手順を示すフローチャート

【図9】図5中の代表音声素片生成部での第3の実施形態による処理手順を示すフローチャート

【図10】本発明の別の実施形態に係る音声合成装置のブロック図

【図11】図10中の代表音声素片生成部での第1の実施形態による処理手順を示すフローチャート

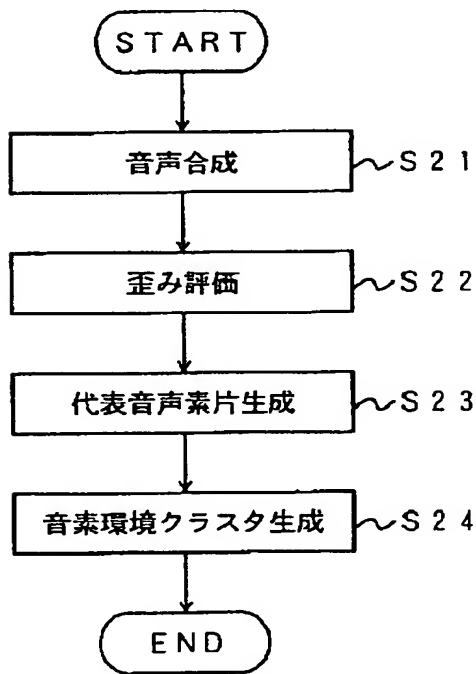
【図12】図10の中の代表音声素片生成部での第2の実施形態による処理手順を示すフローチャート

【図13】図10中の代表音声素片生成部での第3の実施形態による処理手順を示すフローチャート

【符号の説明】

- 1 … 合成単位学習系
- 2 … 規則合成系
- 1 1 … 代表音声素片生成部
- 1 2 … 音素環境クラスタ記憶部
- 1 3 … 代表音声素片記憶部
- 1 4 … 素片選択部
- 1 5 … 音声合成部
- 1 6 … 適応ポストフィルタ
- 2 1 … ホルマント強調フィルタ

【図2】



2 2 … ピッチ強調フィルタ

1 0 1 … トレーニング音声素片 (第1の音声素片)

1 0 2 … トレーニング音声素片にラベル付けされた音素環境

1 0 3 … 入力音声素片 (第2の音声素片)

1 0 4 … 代表音声素片

1 0 5 … 音素環境クラスタ

1 0 6 … 音素環境クラスタ

1 0 7 … 代表音声素片選択情報

10 1 0 8 … 代表音声素片

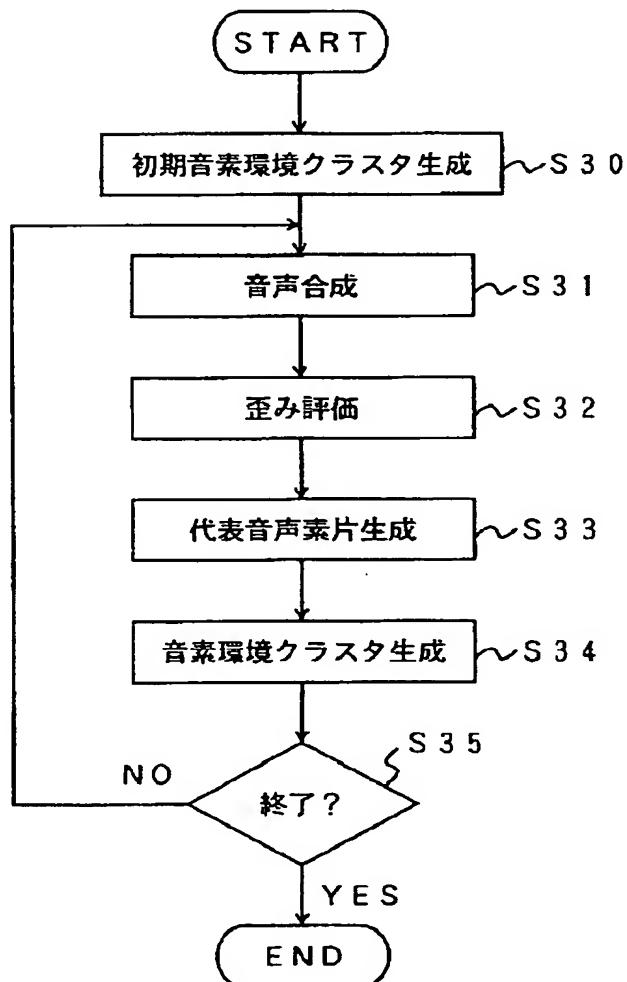
1 1 1 … 韻律情報

1 1 2 … 音韻記号列

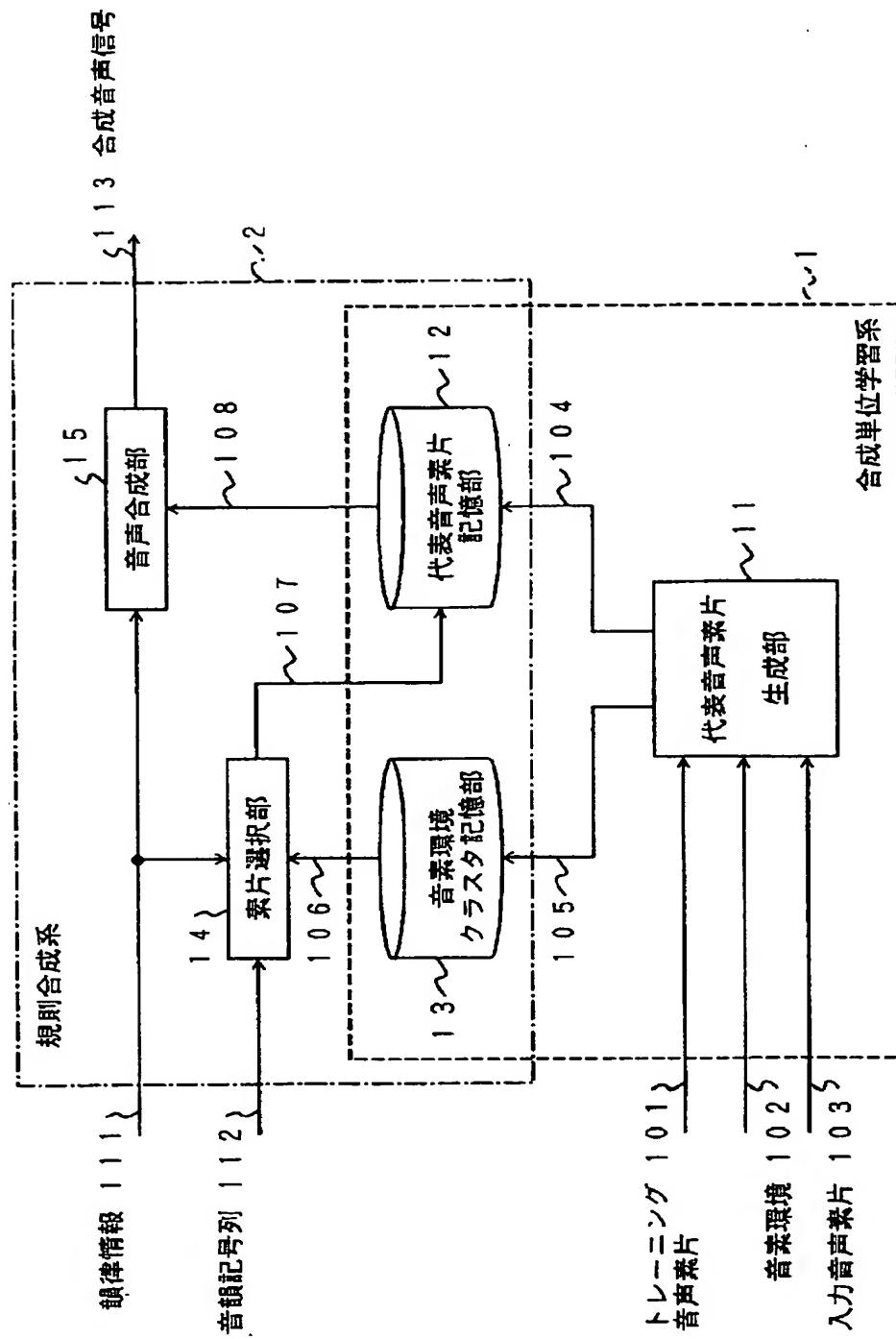
1 1 3 … 合成音声信号

1 1 4 … 合成音声信号

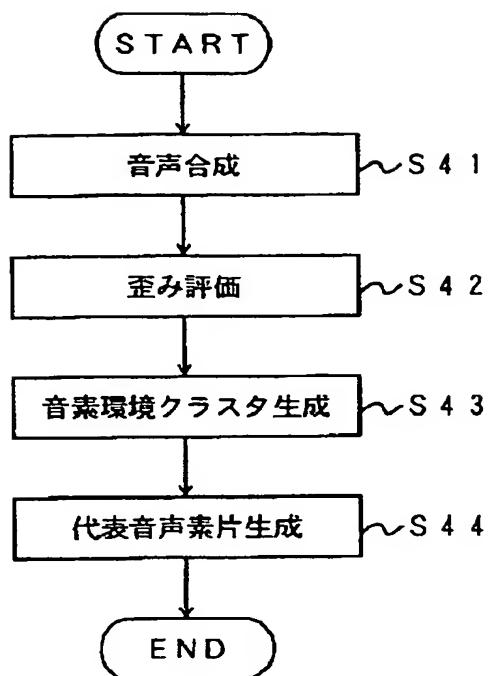
【図3】



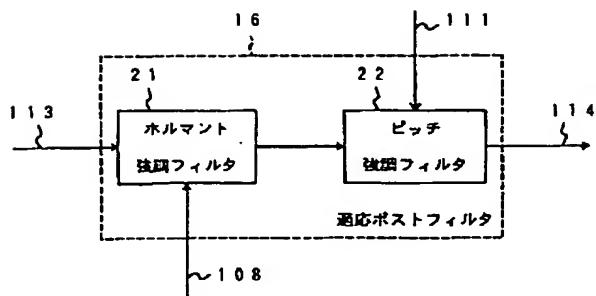
【図1】



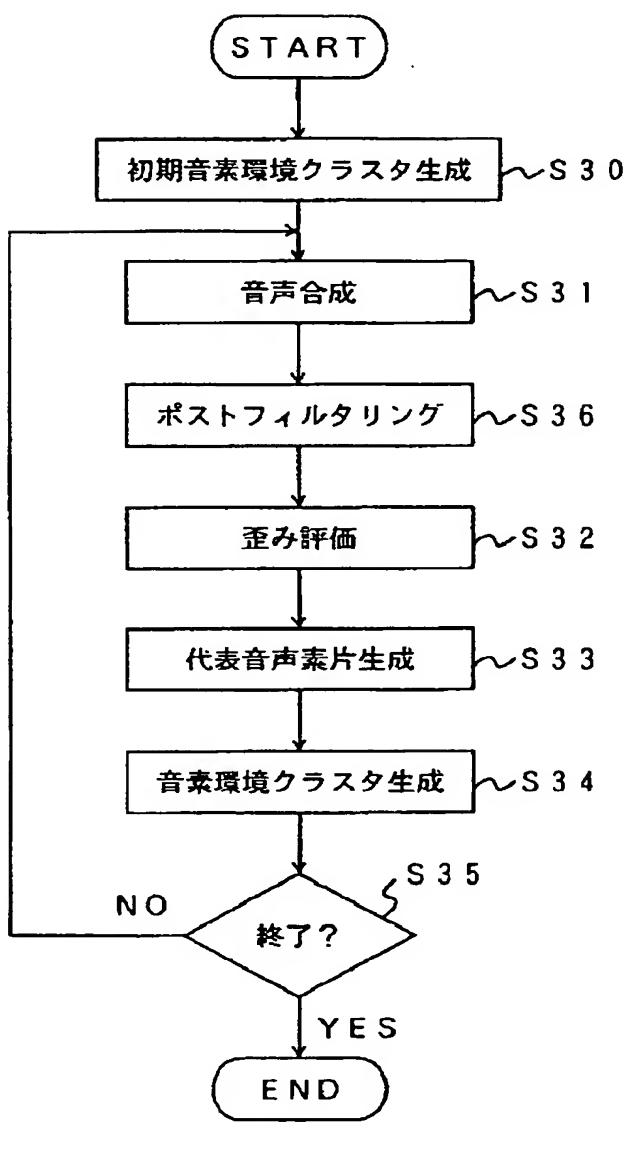
【図4】



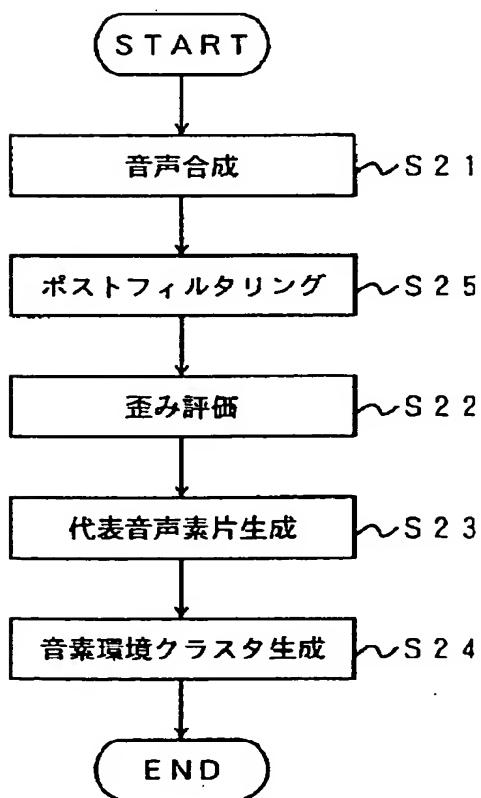
【図6】



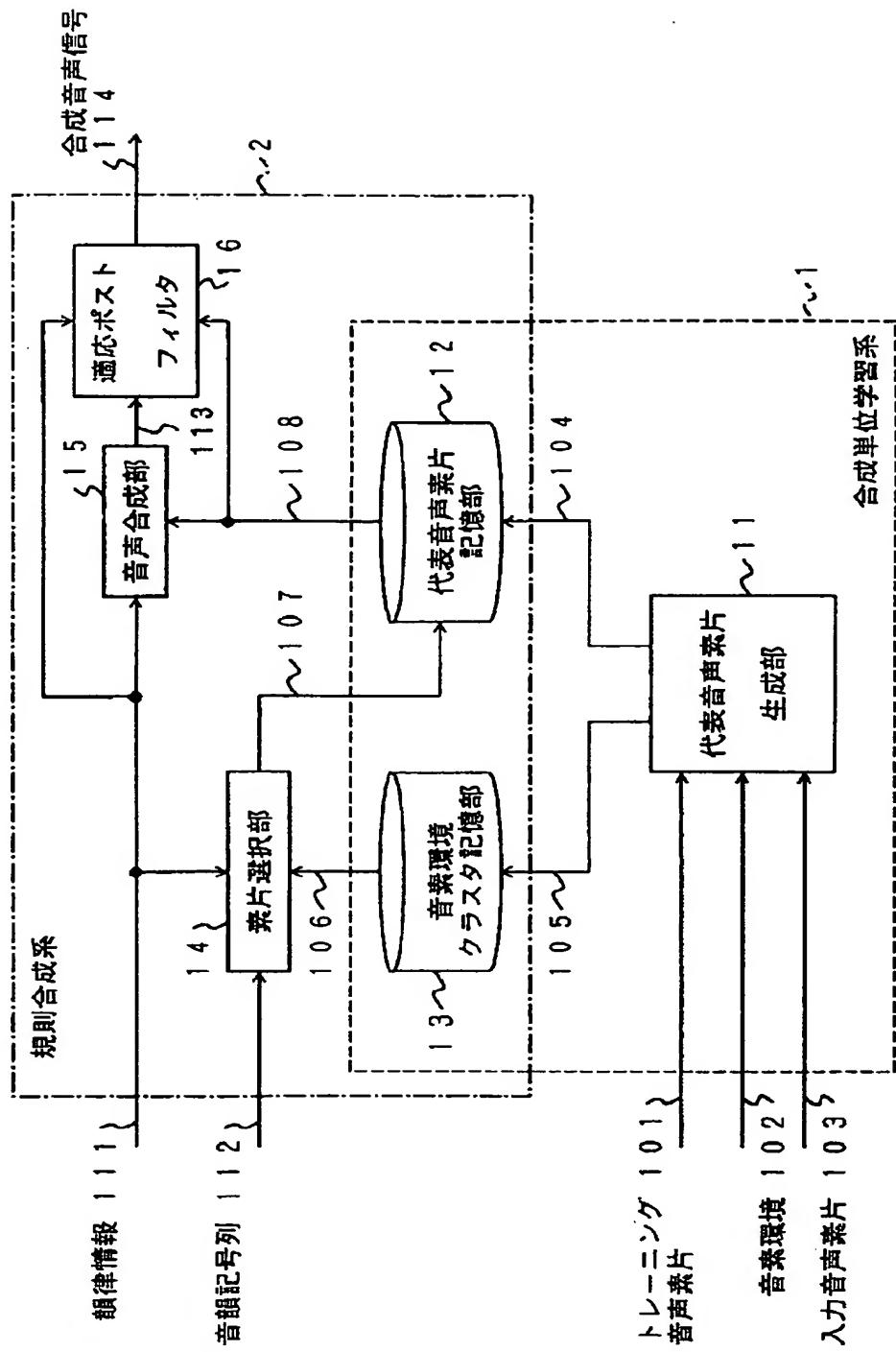
【図8】



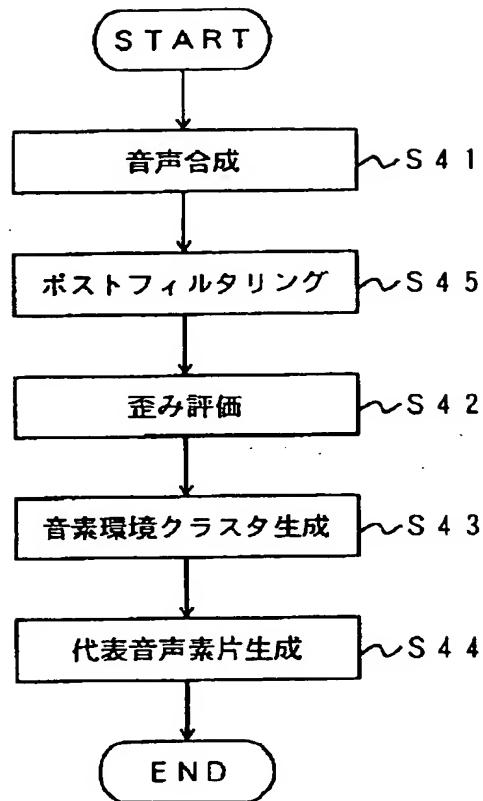
【図7】



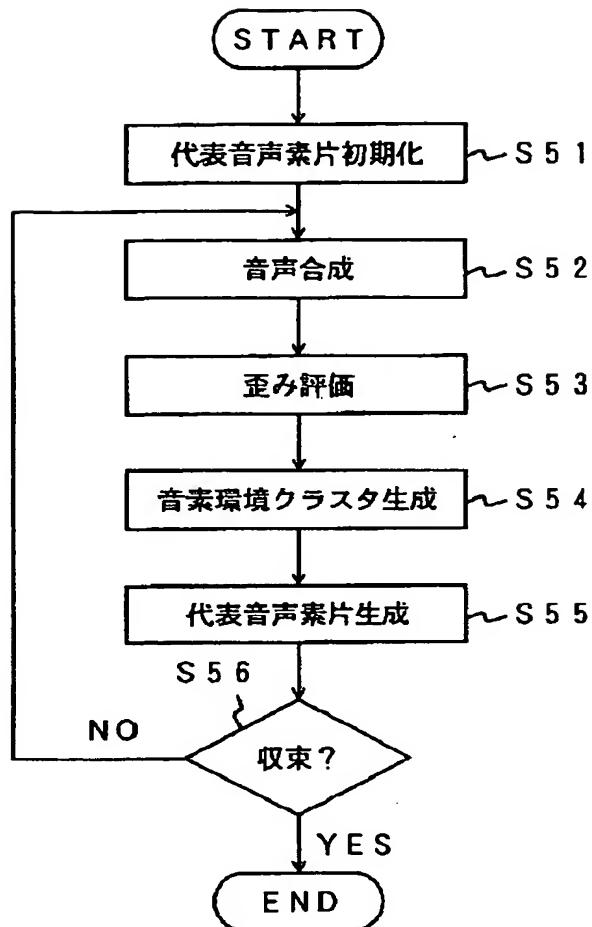
[図 5]



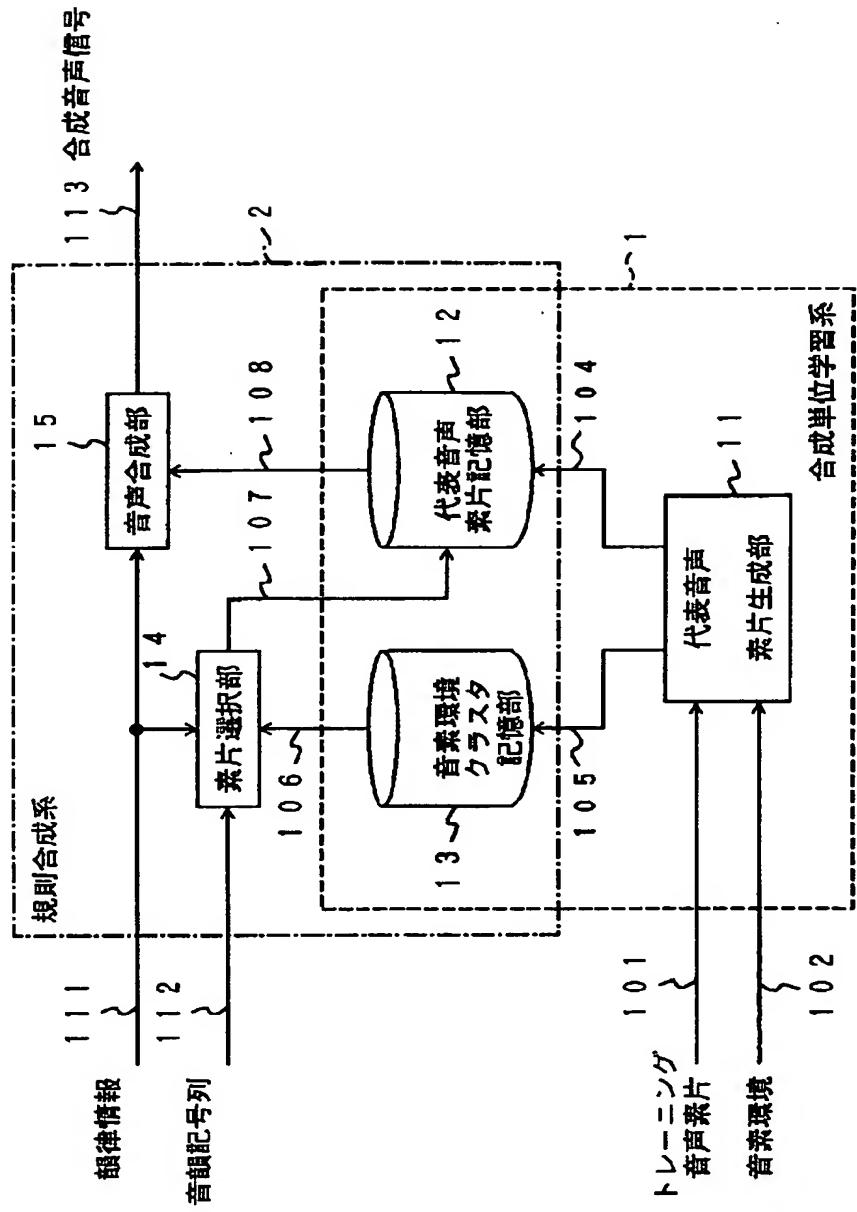
【図9】



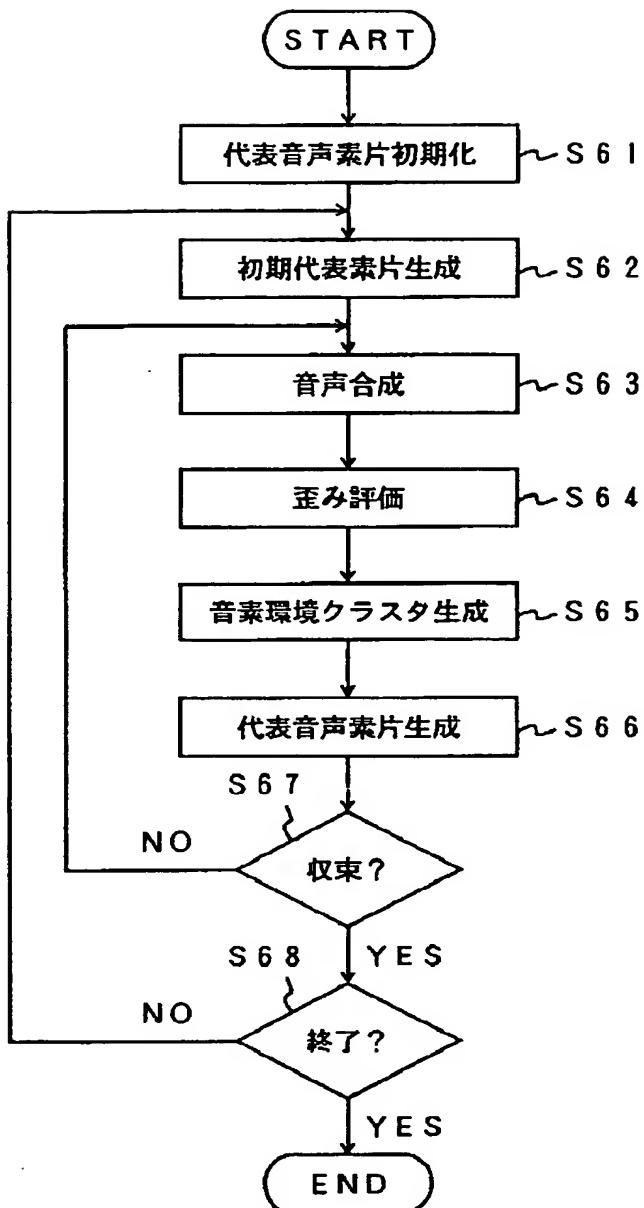
【図11】



【図10】



【図12】



【図13】

